JNL2KRシステムを用いた日本語から意味表現への変換

樫原和昭[1]　植松すみれ[2,4]　宮尾祐介[2,3,4]　Chitta Baral[1]

[1]アリゾナ州立大学　[2]国立情報学研究所　[3]総合研究大学院大学　[4]JST さきがけ

{kkashiha, chitta}@asu.edu, {uematsu, yusuke}@nii.ac.jp

1　はじめに

自然言語を用いてシステムと対話するためには，自然言語の文をシステムの入力言語へと翻訳する必要がある．しかし，システム毎に入力言語 (意味表現) が異なるため，自然言語からアプリケーションが必要とする様々な意味表現へ変換する方法が必要となる．現在，英文から様々なアプリケーションの意味表現へ翻訳するシステムは存在するが，稲田ら [14] が述べているように，日本語文から意味表現へ変換するツールまたはプラットフォームは存在していない．

WASP [12]，λ-WASP [13]，UBL 及び UBL-s [5]，そして FUBL [6] などローマ字表記された日本語音訳から意味表現へ翻訳するシステムは存在するが，日本語音訳は日本語が持つ特徴を捉える事ができない．例えば，“Give me the cities in Virginia” [12] は，日本語訳“バージニアの都市はなんですか” を，形態素解析された日本語音訳で “baajinia no toshi wa nan desu ka” と表記している．“toshi” はこの場合 “都市” であるが，“年” または “歳” と捉えることもできる．つまり，音訳は日本語の持つ情報，特に漢字に関連する情報を失っている．また，この音訳はすでに形態素解析された前処理済みの文であり，これらのシステムは日本語文を入力とする他のシステムへの応用は難しい．

NL2KR [1, 4, 11] は semantic parsing(意味的構文解析) プラットフォームであり，英文とその意味表現の組である訓練例と小規模な初期辞書を用いて半自動で英文から目的の言語 (意味表現) へ翻訳するシステムを構築し，最近の研究 [11] では高い精度を示している．英文を構文解析するため，NL2KR システムは組合せ範疇文法 (Combinatory Categorial Grammar: CCG) パーザを使用している．

そこで本稿では，能地ら [16] の日本語 CCG パーザを利用することで，日本語文から意味表現へ翻訳するシステムを半自動で構築する初のプラットフォーム JNL2KR を提案する．JNL2KR は通常表記の日本語文を入力とするため，日本語を入力とするコマンド解釈や質問応答システムなどに応用できる事が期待される．

我々は JNL2KR を標準的なデータセットである GeoQuery250 [12] を用いて評価した．ただし，日本語はこのコーパス上では日本語音訳であるので，日本語翻訳の専門家によって元の英文から日本語文に翻訳し，その日本語文で JNL2KR を評価した．評価実験の結果は，JNL2KR が極めて小さな初期辞書を用いて最先端の精度を得られる事を示している．

2　研究背景

2.1　NL2KR

NL2KR [1, 4, 11] は，初期辞書や翻訳付きの例文 (意味表現はアプリケーションによって異なる) を用いて，英文から意味表現へ翻訳するシステムを半自動で構築する，使用者にとって使いやすいプラットフォームである．このプラットフォームは，単語と句の持つ意味をラムダ計算表現として表現する Montague の手法 [8] を用いている．ラムダ計算表現では，文の意味は適切なラムダ計算 [2] の関数適用を介して構成単語の意味から構築する．NL2KR は学習モジュールと翻訳モジュールに分かれている．

学習モジュールでは初期辞書と翻訳付きの例文を元に，CCG パーザで例文を解析し，未知の単語の意味を逆ラムダ計算及び Generalization[1] で学習することで，初期辞書よりも多くの単語の意味を収録した辞書を出力する．例えば，“John eats rice eat(john, rice)” という例文と “John N john” (単語: John，CCG カテゴリ: N，意味表現: john) という初期辞書が与えられた時，図 1 に示すように，Generalization で “rice” の意味が “John N john” の例から “rice N rice” と学習され，逆ラムダ計算で “eats rice” 及び “eats” の意味がそれぞ

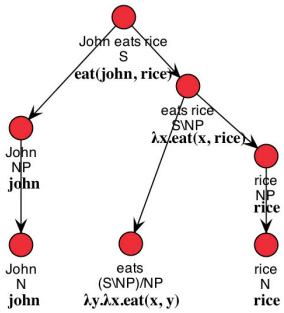

図 1: 英文 “John eats rice” の CCG 解析木と各単語の意味

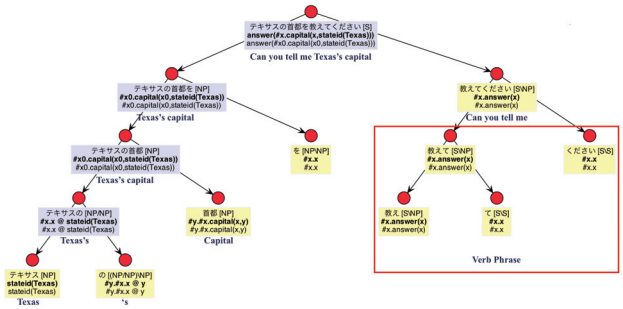

図 2: 日本語文”テキサスの首都を教えてください”の意味表現付き CCG 解析木

れ “λ x.eat(x,rice)” 及び “λ y.λ x.eat(x,y)” と計算され，“John N john”，“eats (S\NP)/NP λ y.λ x.eat(x,y)”，“rice N rice” を収録した辞書が出力される．

翻訳モジュールでは，翻訳したい文をパーザで解析し，解析木と学習モジュールで出力された辞書から文の意味表現を計算する．つまり，NL2KR システムでパーザから精度の高い解析木を得る事は重要であり，JNL2KR も同様に精度の高い解析木を得る日本語 CCG パーザが必要である．

2.2 日本語 CCG パーザ

組合せ範疇文法 (CCG) は統語論の一つであり [9]，CCG での文の導出は文全体の意味を確立するために各単語が結合する方法を決定する．自然言語の文は様々な CCG 解析木を持つことができ，各木は文の異なる意味を表現する．既存の英語 CCG パーザ [3, 7] は，いずれも与えられた文または句の全ての可能性のある解析木の中から，最も確率の高い木を出力する．

日本語では，戸次 [15] によって組合せ範疇文法に基づく日本語構文のための総合的な理論が提案されている．能地ら [16] は植松ら [10] が構築した日本語 CCG バンクを元に，初の日本語 CCG パーザを開発している．

3 提案手法

JNL2KR はこの日本語 CCG パーザで日本語文の解析を行うが，NL2KR をそのまま適用することはできず，日本語と英語の違いを考慮する必要がある．例えば，日本語で時制を表したり形容する表現といった意味情報は助動詞を用いることで補われている．図 2 の枠で，動詞句 “教えてください (CCG カテゴリ: $S\backslash NP$)” は “教え (自動詞，CCG カテゴリ: $S\backslash NP$)”，“て (接続助詞，CCG カテゴリ: $S\backslash S$)”，そして “ください (補助動詞, CCG カテゴリ: $S\backslash S$)” に分割される．この動詞句の意味表現が単純な $\lambda x.x$ であった場合，この句のすべての単語の意味は同じ $\lambda x.x$ となる．このように，有用でない意味を学習したり初期辞書に与えることを避けるため，NL2KR の学習アルゴリズムに従いつつ，日本語 CCG パーザと日本語文法に JNL2KR を適応させるためのフレーズ上書き機能と呼ぶ新たな関数を導入する．

フレーズ上書き機能は 2 通りの使い方があり，自動フレーズ上書きモードと手動フレーズ上書きモードがある．また，JNL2KR は出力された解析木を拡大縮小したりノードを動かすことができるため，複雑な文の構造を視覚的に知る事ができる．

3.1 自動フレーズ上書きモード

自動フレーズ上書きモードは動詞句を検知し，動詞句の CCG カテゴリが $S\backslash NP$ である場合，その動詞句を出力する解析木の動詞句のノードを終端ノードに上書きする．動詞句は CCG カテゴリ，$S\backslash NP$，$S\backslash NP\backslash NP$，$S\backslash S$ などが割り当てられる．カテゴリ $S\backslash NP\backslash NP$ や $S\backslash S$ などは動詞句のみならず他の句や品詞にも割り当てられることがあるので，我々は $S\backslash NP$ をカテゴリとする動詞句のみ自動フレーズ上書きモードに対応させている．例えば，“テキサスの首都を教えてください” をパーザで解析すると，“教えてください　$S\backslash NP$” は動詞句かつ CCG カテゴリが $S\backslash NP$ であるため，“教えてください　$S\backslash NP$” が終端ノードとなり，図 2 の枠で囲まれた部分が削除された解析木が出力される．カテゴリ $S\backslash NP\backslash NP$ や $S\backslash S$ などで上書きしたい動詞句がある場合は，手動フレーズ上書きモードを使用する．

3.2 手動フレーズ上書きモード

手動フレーズ上書きモードはユーザーが上書きしたい名詞句やその他の句を手動フレーズ上書きセットとして受け取り，その句を終端ノードとして上書き処理をする．ユーザは上書きしたい句とその句の CCG カテゴリの組みを含む手動フレーズ上書きセットをパーザに渡すと，手動フレーズ上書きモードを使用したパーザは手動フレーズ上書きセットに含まれる句とその句の CCG カテゴリが一致する非終端ノードが含まれる解析木に対して，その非終端ノードを終端ノードに上書きしたを解析木を出力する．

手動フレーズ上書き機能を使う場合，フレーズ上書きファイルをフレーズ上書きセットとしてパーザに渡すことにより，上書きされた解析木が出力される．例えば，ユーザが名詞句 “何人” (CCG カテゴリは NP) を終端ノードとして扱いたい場合，タブ区切りされた “何人 NP” という組を手動フレーズ上書きセットに加えて，手動フレーズ上書きモードを使用したパーザで文を解析させるだけで良い．“何人” の CCG カテゴリが NP である解析木の非終端ノードを終端ノードに上書きして出力される．

4 評価実験

4.1 コーパス及び実験方法

日本語訳の GeoQuery コーパスを用いて我々の手法を評価する．元の GeoQuery コーパスは，合計 880 の英文がその質問の解をアメリカ合衆国の地理情報データベースから抽出できる意味表現で書かれたそれぞれの意味をペアとして持っている．Wong と Mooney[12] が 250 文を抜き出し日本語音訳に翻訳したコーパスは，WASP システムの評価に使用され，他のシステムの評価にも使用されている．我々はこのコーパスを GeoQuery250 と呼び，250 文の元の英文から日本語へ日本語翻訳家が翻訳したデータを用いる．

我々は，関連研究 [12, 13, 5, 6] との直接比較を行う為，関連研究の実験で用いられた標準的な 10 回交差検定を用いて評価する．適合率，再現率，F1-尺度を報告する．適合率は出力された論理形式が正しい割合，再現率は正しい論理形式が出力された文章の割合，F1-尺度は適合率と再現率の調和平均である．我々は，テスト文の実際の論理形式と我々のシステムの出力が完全一致した場合，翻訳は正しいと判断する．

4.2 初期辞書と実験条件

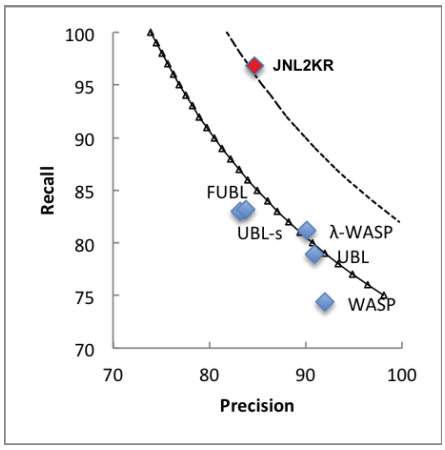

図 3: GeoQuery250 の適合率 (Precision)，再現率 (Recall)，そして，F1-尺度の比較. 破線は F1 値が 90% を示し，実線は F1 値が 85% を示す．

システム (%)	再現率	適合率	F1 値
WASP	74.4	**92.0**	82.9
λ-WASP	81.2	90.1	85.8
UBL	78.9	**90.9**	84.4
UBL-s	83.0	83.2	83.1
FUBL	83.2	83.8	83.5
JNL2KR	**96.8**	84.7	**90.3**

表 1: 日本語 GeoQuery250 データセットの完全一致率

初期辞書を作成するため，我々は 10-検定グループから 2-検定グループをランダムに選び出し，システムが 2 回繰り返し学習するまでに，訓練例文全てを学習するのに必要な意味全てを初期辞書に与えている．

この初期辞書の大きさは，本実験では 250 語である．250 語のうち，31 語が接尾辞，5 語が記号または接続詞，75 語が動詞か助動詞，47 語が形容詞か形容動詞，そして 92 語が名詞または名詞句である．この中には同じ単語で複数の異なる意味を持つ場合も含まれる．

4.3 結果

図 3 及び表 1 は関連研究と GeoQuery250 データセットでの JNL2KR の精度比較を示している．表 1 は，変数なしの意味表現を用いた WASP の精度，ラムダ計算の

意味表現を用いた λ-WASP，UBL，UBL-s 及び FUBL の精度を示す．GeoQuery250 データセットの 7 つの文で日本語 CCG パーザが解析木を出力できず，JNL2KR の適合率の値は WASP，λ-WASP，そして UBL よりも低かった．これは，CCG パーザの曖昧性解消モデルあるいは文法に起因するもので，今後の改善が期待される．WASP，λ-WASP，UBL，UBL-s，及び FUBL は日本語音訳のみ対応しており，すでに単語または句単位で分割されていた日本語音訳では構文解析の問題は少なかったと思われる．他方で，JNL2KR は未知の単語の意味をほぼ学習し，日本語文からその文の意味表現へ 90% 以上の精度で変換できたため，JNL2KR の再現率と F1 値は他のシステムよりも高かった．

5 おわりに

本研究において，日本語文から同等の意味表現へ変換する翻訳システムを構築する為のプラットフォーム JNL2KR[1] を開発した．JNL2KR は自由配布され，GUI を用いるなどユーザに配慮したプラットフォームとなっている．

我々は本論文で JNL2KR のアルゴリズムについて説明し，GeoQuery250 データセットを用いて評価を行った．その結果，JNL2KR が最も高い再現率と F1 値を示した．これは，我々のシステムが効率よく未知の単語の意味を学習し，日本語文から目的の意味表現へ効果的に翻訳を行うことを示している．しかし，我々が使用したパーザは初の日本語 CCG パーザであり，改善の余地がある．このパーザの精度が向上でき，問題のある文を解析できるようになれば，解析木がない，または間違った構造の解析木が出力される問題が解決され，適合率の向上が期待される．

参考文献

[1] C Baral, J Dzifcak, K Kumbhare, and N H Vo. The NL2KR system. In *Proceedings of LPNMR 2013*, 2013.

[2] A Church. An Unsolvable Problem of Elementary Number Theory. *American Journal of Mathematics*, Vol. 58, No. 2, pp. 345–363, April 1936.

[3] J Curran, S Clark, and J Bos. Linguistically Motivated Large-Scale NLP with C&C and Boxer. In *Proceedings of ACL 2007*. ACL, 2007.

[4] S Gaur, N H Vo, K Kashihara, and C Baral. Translating simple legal text to formal representations. In *JURISIN 2014*, 2014.

[5] T Kwiatkowski, L Zettlemoyer, S Goldwater, and M Steedman. Inducing probabilistic CCG grammars from logical form with higher-order unification. In *Proceedings of EMNLP 2010*. ACL, 2010.

[6] T Kwiatkowski, L Zettlemoyer, S Goldwater, and M Steedman. Lexical generalization in ccg grammar induction for semantic parsing. In *Proceedings of EMNLP 2011*. ACL, 2011.

[7] Y Lierler and P Schüller. Parsing combinatory categorial grammar via planning in answer set programming. In *Correct Reasoning*, pp. 436–453. Springer, 2012.

[8] R Montague. English as a Formal Language. In *Formal Philosophy: Selected Papers of Richard Montague*, pp. 188–222. Yale University Press, 1974.

[9] M Steedman. *The Syntactic Process*. MIT Press, Cambridge, MA, 2000.

[10] S Uematsu, T Matsuzaki, H Hanaoka, Y Miyao, and H Mima. Integrating multiple dependency corpora for inducing wide-coverage japanese CCG resources. In *Proceedings of ACL 2013*, 2013.

[11] N H Vo, A Mitra, and C Baral. The nl2kr platform for building natural language translation systems. In *Proceedings of ACL*, 2015.

[12] Y W Wong and R J. Mooney. Learning for semantic parsing with statistical machine translation. In *Proceedings of HLT-NAACL 2006*, 2006.

[13] Y W Wong and R J. Mooney. Learning synchronous grammars for semantic parsing with lambda calculus. In *Proceedings of ACL 2007*, 2007.

[14] 稲田和明, 松林優一郎, 井之上直也, 乾健太郎. 効率的な推論処理のための日本語文の論理式変換に向けて. 言語処理学会第 19 回年次大会, 2013.

[15] 戸次大介. 日本語文法の形式理論. くろしお出版, 2010.

[16] 能地宏, 榊原隆文, 宮尾祐介. 日本語パイプライン処理のための簡易フレームワークの提案. 言語処理学会 第 21 回年次大会, 2015.

[1] https://goo.gl/WR8xs6